PAC267-C
ESX Server Storage Internals

Satyam Vaghani
Staff Engineer, R&D

19 October 2005
Outline

- ESX Server storage architecture
- Storage virtualization core
- Virtual machine state management
- Managing physical storage devices
- Questions
Outline

- ESX Server storage architecture
- Storage virtualization core
- Virtual machine state management
- Managing physical storage devices
- Questions
Goals for ESX Server Storage

- High speed, isolated access from virtual machine to its disk
- Organize storage hardware into a distributed, structured resource pool
- Hide physical storage complexity from virtual machines
- High availability, scalability, and reliability
- Lower cost of data protection for virtual machines
ESX Server Storage Architecture

- **Service console (SC)**
  - VMFS driver exports VMFS volumes to /vmfs, forwards VMFS file access requests to VMKernel FS core
  - VMKDev driver exports all VMKernel storage devices to SC kernel, forwards requests to storage core

- **Virtual machine**
  - Uses file-level functions to setup and manage virtual disk files
  - Uses SCSI commands to do IO from Guest OS to virtual SCSI device

- **VMKernel**
  - Provides buffered IO interface for POSIX-style file requests
  - Virtual disk IO goes directly to FS or storage core

[Diagram showing the architecture with nodes for Service Console, VMFS, VMKDev, VMKernel, Virtual Machine, Virtual Disk Setup, Virtual HBA, Data Cache, Virtual Storage, Virtual Machine File System, and Storage Core (Physical Device Access).]
ESX Server Storage Stack

- Virtual Machine
- SCSI Virtualization Engine
- VMFS
- Logical Volume Manager
- Storage Core
- Multipathing
- Device Driver

**Flow**: SCSI command → FS operation → Block operation → SCSI command → To Storage Device

- Storage Virtualization Core
- Virtual Machine State Management
- Physical Device Access
Outline

- ESX Server storage architecture
- **Storage virtualization core**
- Virtual machine state management
- Managing physical storage devices
- Questions
SCSI Virtualization Engine

- Exports a file, RDM, LUN or redo log as virtual SCSI disk
- Forwards, filters or remaps commands (downstream) and IO completions (upstream)
- New in ESX Server3
  - Hot-add virtual disks to virtual machines
  - Layered apps inside virtual machines with RDMs
  - Uniform virtual disk management across VMware products

NOTE: ESX Server3 features still under development and may be subject to change
Virtual SCSI Devices

SCSI Commands From Virtual Machine

- SCSI-2 compliant VMware virtual disk
- Physical device access (pass-through)

Device State

- Test Unit Ready, Inquiry
- Reserve, Release, Reset
- Read, Write
- Read Buffer, Write buffer
- Get Capacity
- Busy
- Filter
- Filter
- Report LUNs
- SCSI Command

SCSI Virtualization Engine

To/From File System

- Virtual Resv, Rel, Rst
- File Read, File Write
- File Stat
- No Connect

To/From Storage Core

- SCSI Command
## Virtual Disks

<table>
<thead>
<tr>
<th>Type</th>
<th>Format</th>
<th>Performance</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preallocated</td>
<td>Flat</td>
<td>High</td>
<td>ESX Server default</td>
</tr>
<tr>
<td>Allocate-on-demand*</td>
<td>Flat</td>
<td>Medium to high</td>
<td>Storage over-commitment</td>
</tr>
<tr>
<td>Delta (redo log)</td>
<td>Sparse</td>
<td>Medium</td>
<td>Virtual machine snapshots, backup, DR</td>
</tr>
<tr>
<td>RDM</td>
<td>N/A</td>
<td>High</td>
<td>Layered apps, clustering</td>
</tr>
<tr>
<td>Raw/System LUN</td>
<td>N/A</td>
<td>High</td>
<td>Deprecated, expose non-disk devices to virtual machine</td>
</tr>
</tbody>
</table>

* New in ESX Server3

NOTE: ESX Server3 features still under development and may be subject to change
Outline

- ESX Server storage architecture
- Storage virtualization core
- Virtual machine state management
- Managing physical storage devices
- Questions
The ESX Server Virtual SAN

Virtual disks are LUNs on the “VMFS Disk Array”

Legend
- VMFS-3 volume
- Base disk
- Redo log (Snapshot)
- RDM
- Raw Device

Virtual Machine 1
Virtual Machine 2
Virtual Machine 3
Virtual Machine 4

Storage Virtualization Device (/vmfs)

VMFS vol 1
1.vmdk
1.vmdk.redo
2.vmdk

VMFS vol 2
3.vmdk
2.vmdk.redo
4.rdm

Virtual Disk Array 1
Virtual Disk Array 2

vmhba0:0:0:0
Virtual Machine File System (VMFS)

Optimized for accessing large files from VMM
- Keeps virtual disk performance close to native

![Graph showing performance comparison between different block sizes and operations.](image-url)
VMFS, cont.

- Enhanced functionality on SANs
  - Distributed access from ESX Server hosts
  - No network lock manager, or knowledge of other hosts
  - Auto-discover volumes, SAN volume manager, RDMs, snapshots
- Virtual machine storage consistency
  - VMFS partition protection, exclusive locks across hosts, crash consistent virtual machine IO path
- Special primitives for clustered virtual machines, raw LUNs
- Easily manage virtual disks as files
  - Enables elegant backup and DR solutions
VMFS-3
A New FS For ESX Server3

- Exclusive repository of virtual machines and virtual machine state
  - Better organization through directories, small files
  - Large number of files to host more virtual machines
  - Enables DAS and DRS
- Stronger consistency mechanisms
  - Distributed journal for faster crash recovery
  - Crash recovery and metadata update code is tested in normal IO paths

NOTE: ESX Server3 features still under development and may be subject to change
VMFS-3
Performance

- Reduced IO-to-disk for metadata operations
- Less contention on global resources
- Less disruption due to SCSI reservations
- Faster virtual machine management operations

NOTE: ESX Server3 features still under development and may be subject to change
VMFS-3
Scalability

- Large number of FS objects don’t compromise performance
- Greater connectivity (hosts or virtual machines/VMFS volume)
- Fairness across multiple virtual machines hosted on the same volume

NOTE: ESX Server3 features still under development and may be subject to change
Logical Volume Manager

- Consolidates multiple physical disks into a single logical device
- New in ESX Server3
  - Volume availability not compromised due to missing disks
  - Automatic resignaturing for volumes hosted on SAN snapshots

NOTE: ESX Server3 features still under development and may be subject to change.
Network File System (NFS)

- New in ESX Server 3: NFS v3/TCP driver to mount NAS exports in VMKernel
  - Cheaper shared storage alternative to SAN
  - Easier to provision and setup
    - No need to carve out LUNs, plug in HBAs, set multipathing policies
    - Leverage existing IP networks
  - Lower performance than FS on DAS/FC-SAN
  - Plugs into ESX Server storage stack as a FS driver similar to VMFS-3

NOTE: ESX Server 3 features still under development and may be subject to change
Outline

- ESX Server storage architecture
- Storage virtualization core
- Virtual machine state management
- Managing physical storage devices
- Questions
Storage Core

- Manages access to physical SCSI devices
- Command queuing and disk scheduling
- PAE copies, command splitting based on hardware capabilities
- Handles aborts and retries for system IO
Native multipathing (MP) driver

- Highly available connectivity to FC-SAN storage
  - Path scanning to discover all paths to a SCSI device
  - Path collapsing to export the discovered device to upper layers
  - IO scheduling policy for active paths
  - Transparent path failover
- New in ESX Server3
  - MP starts at boot time for boot-from-SAN
  - Supports new disk arrays and storage virtualization devices

NOTE: ESX Server3 features still under development and may be subject to change
Device Management

- Dynamically add/remove storage devices from ESX Server
- VMFS information is automatically updated
- New in ESX Server3
  - All devices show up in VMKernel Device File System (DevFS)
  - Management functions like fdisk, cp can be performed directly on DevFS nodes
  - UUIDs as persistent names for disks and VMFS volumes. All devices have user-friendly names too

NOTE: ESX Server3 features still under development and may be subject to change
iSCSI

- New in ESX Server3: Access to iSCSI storage devices through VMKernel
  - Leverage existing IP networks for storage traffic
  - Suitable for smaller installations, tier 2/3 workloads
  - Support for both s/w and h/w iSCSI initiators
  - Plugs into ESX Server Storage Stack as device driver similar to other SCSI and FC drivers
    - Implicitly suitable for VMFS-3, RDM and raw device access

NOTE: ESX Server3 features still under development and may be subject to change.
Summary

- ESX Server contains a high performance, reliable and scalable storage stack for accessing and managing virtual machine state
  - Simplifies advanced datacenter technologies for virtual machines
  - Lowers the cost of managing virtual machine state
  - Enables low-cost and elegant disaster recovery and data protection solutions
Questions